Active Learning Driven PyTorch Framework for Automated Chest X-ray Labeling and Disease Classification Using NIH Dataset

INTRO/ABSTRACT

This project develops an **Active Learning System for chest X-ray image classification**, designed to automate data preparation, streamline model training, and reduce manual labeling effort. It provides a structed, scalable pipeline from raw data ingestion to automated decision-making, including dataset indexing, patient-aware splitting, preprocessing, configuration management, and validation.

The model uses **CNNs** to generate baseline diagnostic predictions across chest pathologies. Active learning strategies then identify the most informative unlabeled images for iterative retraining, improving performance while minimizing labeling cost.

METHODS

PyTorch provides an easy and flexible way to build and train deep learning models while also having wide community adoption.

Kaggle provides the **NIH Chest X-rays** dataset that our model uses for training, testing, and validation.

GitHub provides version control, allows collaboration on project files, and hosts the project website with **GitHub Pages**

RESULTS

We are currently in the model training and validation phase of development. Our work to date focuses on establishing a baseline classification model that will serve as the foundation for the subsequent workflow.

Proposed strategies for training improvements include transfer learning with a pretrained model. Evaluation tools such as a dashboard to track performance would also enable a reproducible, clinically relevant image workflow.

Deep Learning Example Diagram

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Authors – Josh Smith, Elijah Merrill, Noah Lane, Matthew Hall
Advisor – Sharon Perry